Updated Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies

Kirill Semenov[C] Vilém Zouhar[E] Tom Kocmi[M] Dongdong Zhang[M]
Wangchunshu Zhou[A] Yuchen Eleanor Jiang[A]

[C]**Charles University** [E]**ETH Zürich** [M]**Microsoft** [A]**AIWaves**

NB: This is the updated version of the document, as after the publication of WMT23 proceedings, several inconsistencies in authors' submissions and our metric implementation were found. The updates cover the content of tables 5-8, as well as the choice of success rate metric (Section 5). Release date: 16.12.2023.

Abstract

The WMT 2023 Terminology Shared Task investigates progress in machine translation of texts with specialized vocabulary. The participants were given the source text and segment-level terminology dictionaries for three language pairs: Chinese→English, English→Czech, and German→ English. We evaluate 21 submissions from 7 teams on two main criteria: general translation quality and the effectiveness of translating specialized terminology. Systems took varied approaches — incorporating terminology at inference time or weakly supervised training that uses terminology access. While incorporating terminology dictionaries leads to improvement in the translation quality, incorporating an equal amount of information from the reference leads to similar results. This challenges the position of terminologies being the crux of meaning in translation, it can also be explained by inadequate metrics which are not terminology-centric.

1 Introduction

General-purpose machine translation models often show limitations when applied to specialized tasks, like translating specialized vocabulary. This gap is critical in medicine, science, and law, where language precision is paramount — medical inaccuracies, juridical misunderstandings, and technological malfunctions can lead to serious problems. The translation of technical terms is not a mere exercise in lexical fidelity — it supports effective communication in highly specialized fields. Terminology correctness and consistency has already been long in focus from the modelling (Dinu et al., 2019; Hasler et al., 2018), evaluation (Zouhar et al., 2020; ibn Alam et al., 2021; Semenov and Bojar, 2022) and translators' perspective (Cabré, 2010; Vargas-Sierra, 2011; Arcan et al., 2017).

Source	Der Bericht entspricht FOG.
Reference	The report is ROA-compliant.
Hyp. 1	The report is in accordance with FOG.
Hint 1	"FOG" → "ROA"
Hyp. 2	The report is in accordance with ROA.
Hint 2	"entspricht" → "compliant"
Hyp. 3	The report is compliant with ROA.

Table 1: Translation with "terminologies". Hyp. 1 is without any hints and the worst while Hyp. 3 is close to the reference. Hint 1 is proper terminology while Hint 2 only helps align the translation with the reference. Does terminology-assisted MT work because of Hint 1 or because it leaks information from the reference?

We shed light into recent advancement in this area by assessing MT systems with segment-level terminology dictionaries. Alongside the general evaluation of translation quality, our shared task emphasizes the *effectiveness* terminology dictionaries. This task follows the latest efforts on evaluating progress in terminology-enhanced translation (Alam et al., 2021). While we are also concerned with the quality of the translation, we refocus on measuring the relative improvement of incorporating the terminology dictionary. Focusing on System A being overall better with terminologies than System B might obscure the fact that System A is already good without terminologies while the methods of System B improves. From research perspective, System B gives us more insight into how to more efficiently incorporate terminology dictionaries. Additionally, it disentangles the terminology-incorporation meth-

Perf. ↑	**A**	**B**
Base	95	90
+Dict.	92	70

German→ English	**Source**: “Most informative is the analysis of airway secretions:” **Reference**: “Häufig jedoch führt die Analyse von Material aus den Atemwegen zur Diagnose:” **Proper**: “analysis of airway secretions” →“Analyse von Material aus den Atemwegen” **Random**: “Most”→“Häufig”
English→ Czech	**Source**: “We present Eman, an experiment manager, and show how to use it to train several simple MT systems.” **Reference**: “Popisujeme Emana, nástroj pro správu experimentů, a ukazujeme, jak ho lze využít k trénování několika jednoduchých systémů pro strojový překlad.” **Proper**: “Eman”→“Emana”, “an experiment manager”→“nástroj pro správu experimentů”, “MT systems”→“systémů pro strojový překlad” **Random**: “how to use”→“jak ho lze využít”, “train”→“trénování”, “simple”→“jednoduchých”
Chinese→ English	**Source**: “凌寒再次挥手，又结结实实地抽了他一巴掌。” **Reference**: Ling Han raised his hand once more, and gave him another solid slap. **Proper**: ‘凌寒”→“Ling Han” **Random**: ‘手”→“his hand”

Table 2: Examples from the WMT 2023 Terminology Shared Task test dataset, based on MuchMore Springer Bilingual Corpus, Rosa and Zouhar (2022), and Jiang et al. (2023). **Base** is without any terminologies, **Proper** is real terminologies and **Random** are random but aligned phrases from source to the reference.

ods from the general MT methods.

This shared task provides one repackaged and two newly-annotated datasets which can be used for segment-level terminology enhanced machine translation evaluation.[1]

2 Task Description

We focus on how translation quality improves with the incorporation of segment-level terminology on German→English, English→Czech, and Chinese→English datasets. Participants are given source sentences along with a segment-level terminology dictionary (*Source* and *Hints* in Table 1). For the purposes of this study, we define terminology as low-frequency words or phrases that occur typically within a particular domain, such as computer science paper abstract. We scan the source and references for such phrases and provide this segment-level annotation, together with the source, to the participants in the form $X \rightarrow Y$ where X is a span from the source and Y is a span from the reference (*Proper* in Table 2).

Given that the participants are given a part of the reference, Y, this raises the following question: *Is the improvement in translation quality due to the information that a particular terminology X is translated as Y or merely because a part of the reference is leaked to the model?* To better attribute any performance gains, we therefore also test a different mode, where we give the participants “terminologies” where X' and Y' are still aligned spans and translations of each other, but sampled randomly. That is, they are treated as terminology but are, in fact, random phrases (*Random* in Table 2). For this reason, we ask the participants to carry out the translation in three distinct modes:

- **Base**: MT with no terminology dictionary.
- **Proper**: MT with a terminology dictionary. For example *“Sprachmodell” → “language model”*.
- **Random**: MT with randomly chosen, but correct, non-terminological translations. For example *“Hund” → “dog”*.

By comparing performance across these modes, we isolate the model’s inherent translation ability and its ability to make use of the terminology.

3 Data

For MT training, the participants were restricted to only the parallel or monolingual datasets enumerated in the WMT general track (Kocmi et al., 2023).[2] The inclusion of pre-trained models was permitted, provided that such usage was explicitly declared. Any employment of terminology-specific datasets that were not part of the specified resources was expressly disallowed. For the terminology-targeted evaluation, we repurposed one dataset and

[1] Public terminology datasets Chinese→English (repack), English→Czech (new data) and German→English (new data): github.com/wmt-terminology-task/data-2023

[2] www2.statmt.org/wmt23/translation-task.html

created two new ones. From all of them, we provided 100 segments to the participants as a sanity-check development set. See examples for all language pairs in Table 2.

X→Y	Count	X/Y Words	Terms
German→English	2963	22.2/22.6	3.8
English→Czech	3005	25.6/21.6	3.6
Chinese→English	2640	9.7/36.9	1.1

Table 3: Our test dataset size, average number of words per line and average number of terms per segment (equal between *Proper* and *Random*).

3.1 Chinese→English Test Data

Our Chinese→English translation test data is sourced from the BWB corpus (Jiang et al., 2023), which covers web novels annotated with, among others, terminologies. The BWB corpus comprises ~3k sentences across six web novels. These annotations identify each named entity and concept in the sentences, highlighting their co-referred expressions. The average terminology count per line is 1.1 (Table 3). Examples of such terminology are in Table 2. Terminology often faces issues of mistranslation or contextually inconsistent translation. Additionally, MT quality declines when terminology is positioned as the subject due to the Chinese's subject-dropping nature.

3.2 English→Czech and German→English Test Data

For the next two language directions we created a new semi-automatically annotated corpus of terminologies. For English→Czech we used 3k sentence pairs from a dataset of NLP papers abstracts (Rosa and Zouhar, 2022). For German→English we used 3k sentence pairs from a dataset of medical paper abstracts (MuchMore Springer Bilingual Corpus). In both cases, the focus on academic texts was guided by the high occurrence of terminology in this domain (3.8 and 3.6, Table 3).

Term.	Prompt
Proper	Identify and annotate all terminology entities (consider only consecutive words) from source sentence and match them with the counterpart in the translated sentence.
Random	Identify and annotate as many as possible aligned words (consider only consecutive words) between source sentence and the translated sentence.

Prompt 1: The upper prompt formulation extracts proper terminology and the bottom extracts random terminology. See Prompt 2 (Appendix) for the full example with few-shot examples.

Automatic alignment tools usually have lower precision than linguists and linguists have lower recall and the collection is both time and budget consuming. Therefore, to extract the aligned terminology, we use human-machine collaboration. First, we use GPT-4 (OpenAI, 2023) to create aligned terminology pairs from source and references. We use two few-shot prompts to collect the raw alignments (Prompts 1 and 2). Then we ask linguists to validate these alignments and fix those that are incorrect (either missing terminology, wrong alignment or pairs that are not a terminology). For the Czech-English language pair, humans revised approximately 8% of GPT annotations. There is no modification to terminology in the German-English GPT annotations. Consultation with German linguist affirmed that no adjustments were necessary. Nonetheless, further examination is needed to fully assess GPT's proficiency in terminology alignment for German. This task was sponsored by Microsoft and we release both the pre- and post-alignment data for the further research of GPT capabilities.

4 Participants and System Descriptions

We received a total of 21 per-language submissions from 7 teams. We provide short descriptions of their systems, based on the submitted details.

AdaptTerm (Moslem et al., 2023b). The terminology-enriched MT system builds on Moslem et al. (2023a); Haque et al. (2020). It consists of:

1. using an LLM to generate bilingual synthetic data based on the provided terminology;
2. fine-tuning a generic model, OPUS, with a mix of the terminology-based synthetic data generated by #1 and a randomly sampled portion of the original generic data; and
3. generating translations with the fine-tuned model from #2, and then fixing translations that do not include the required terms with an LLM.

Lingua Custodia (Liu, 2023). This submission includes all three language directions. They use two strategies to extract synthetic terminology from the training data. The first one relies on the invariable n-grams between the source and the target sentence, while the second one extracts parallel sentences that appear inside another training sample as one terminology item. Then, they train a Transformer-based model with annotated data using the extracted terminology, identical to Alam et al. (2021). In addition, after the text annotation, they further apply several annotated data filters to reduce some bias introduced by the automatic annotation. The final trained model can be used directly to translate a text with any new terminology.

OPUSCAT (Nieminen, 2023). A standard Transformer system is finetuned with parallel data where parts of the source sentences have been annotated with their corresponding translations in the target sentences, causing the system to learn to copy the annotated target parts from the source sentence into the target sentence. The translations are generated using a series of models, with different fine-tuned terminology models acting as backoff models to the base transformer model, in cases where the base transformer output does not contain the specified terminology.

UEDIN (Bogoychev and Chen, 2023). Their primary system, *twoshot*, is 2-shot decoding where we enforce terminology constraints via terminology hints in the source and if this does not work we use alignment-based methods to identify the mistranslated terminology word on the target side and penalize it, giving the decoder a chance to generate the hinted word. System *Tag* is decoding with terminology hints while *LLM* is an unconstrained contrastive system.

BJTU-LB (no description paper). They train the in-context learning ability of the model, and then concatenate the term translation pairs in front of the sentence to be translated as the context. The model can generate different translation results according to different contexts.

VARCO-MT (Park et al., 2023). The *ForceGen* is a Transformer-based model that is tailored to ensure the appearance of given terminology in the generated output. By modifying the input format and decoding process, it incorporates a copy mechanism on the source side, allowing it to copy the target terminology from the provided terminology pairs. During the generation process, it uses a force decoding technique, which compels the model to actively generate the target terminology as needed. The *TSSNMT* is a novel Transformer-based NMT model that uses a shared encoder to process both input text and terminology. The model then employs cross-attention mechanisms between the two encoder hidden states and passes them through a gate, enabling the model to autonomously decide which pieces of information (input or terminology) to focus on during translation.

The TSSNMT submission files comprised mostly of the sentences unseen in the ground truth files, thus we took into account only subset of the sentences which was the intersection with the ground truth files (57 sentences). For completeness, we include the results in the analysis sections in gray (VARCO-MT$_{\text{TSSNMT}}$), but we urge the readers not to draw any comparisons to other submissions. It is especially important for the consistency and term success metrics, as they are sensitive to the size of the submitted file, thus it is easier to get higher scores on a smaller document.

Huawei. Did not submit system description. The translations are also on a subset not used for final evaluation. We include the results in the analysis sections in gray (Huawei) for completeness but urge the readers not to draw any comparisons to other systems.

5 Evaluation

Our evaluation is focused on: (1) general translation quality, (2) quality of translation of specific terminologies, and (3) efficiency in using segment-level terminology dictionaries.

Standard Metrics. Following recent trends in MT evaluation (Kocmi et al., 2021), we use ChrF (Popović, 2015) and COMET (Rei et al., 2020) for the general translation quality evaluation.[3] While the latter one is generally touted as more robust and correlated more with human judgement, in this case we are also concerned in exact match of n-grams, which is captured by ChrF.

Term Success Rate. In the terminology success rate we compare the machine-translated terms with their dictionary equivalents. Since we have the reference proper and random terminology translations for each sentence, we apply the substring search of these terms to the outputs. We do that with regular expressions and count the matches of the regex search. To minimize the sensitivity to the grammatical inflections in the target sentence, we lemmatize

[3] ChrF uses the defaults from sacreBLEU (Post, 2018) and COMET is `wmt22-comet-da`.

the term phrases and the output sentences before running regex search.

Term Consistency. This metric looks at whether technical terms are translated uniformly across the entire text corpus. We aim for high consistency, measured by the low occurrence of multiple translations for the same term within the text. We use the approach suggested by Semenov and Bojar (2022). Given the source sentences, outputs, and source terms assigned to each sentence, we firstly make word alignment for the source sentences and outputs, and extract the aligned translated terms for each source term occurrence. Then, we automatically choose the "pseudo-reference" terminology translations, based on which translation of which source term occurred in the text first. In the last step, we compare two sets –the real outputs and the pseudo-references for each term occurrence– by means of F_1 score on a scale of 0 (no consistent terminology) to 1 (all terminology translated consistently).

System	ChrF De→En	ChrF En→Cs	ChrF Zh→En
AdaptTerm	61.0	64.4	37.5
Lingua Custodia	61.8	67.7	32.6
OPUS-CAT	68.3⋆	75.1⋆	27.7
$\text{UEDIN}_{\text{LLM}}$	60.0	64.8	41.2
$\text{UEDIN}_{\text{Tag}}$	58.3	64.7	41.0
$\text{UEDIN}_{\text{Twoshot}}$	60.5	62.4	34.5
BJTU-LB			43.8⋆
$\text{VARCO-MT}_{\text{TSSNMT}}$			43.0
$\text{VARCO-MT}_{\text{ForceGen}}$			40.5
Huawei	62.1	58.2	36.8

System	$\text{COMET}_{22}^{\text{DA}}$ De→En	$\text{COMET}_{22}^{\text{DA}}$ En→Cs	$\text{COMET}_{22}^{\text{DA}}$ Zh→En
AdaptTerm	0.801	0.841	0.688
Lingua Custodia	0.735	0.834	0.609
OPUS-CAT	0.828⋆	0.889⋆	0.557
$\text{UEDIN}_{\text{LLM}}$	0.813	0.869	0.757⋆
$\text{UEDIN}_{\text{Tag}}$	0.809	0.868	0.757⋆
$\text{UEDIN}_{\text{Twoshot}}$	0.792	0.835	0.650
BJTU-LB			0.751
$\text{VARCO-MT}_{\text{TSSNMT}}$			0.755
$\text{VARCO-MT}_{\text{ForceGen}}$			0.715
Huawei	0.843	0.887	0.666

Table 4: Averages of ChrF and COMET scores with *Proper* terminology dictionaries. The ⋆ marks best within each column (language) and metric.

5.1 Main Results (Table 4)

We begin the comparison using two standard metrics of MT quality in the case where *Proper* terminology dictionaries were provided. The choice of the best-performing system diverges based on the two metrics: *Lingua Custodia* is selected as the best by ChrF in two language directions, it ranks the same system on Zh→En as the second lowest-performing one. In contrast, COMET ranks $UEDIN_{LLM}$ as the best across all three language directions. Given that this metric better captures human judgement (Freitag et al., 2022), this ranking is likely more close to the true quality.

5.2 Terminology Quality (Table 6)

The results are even more different when focusing solely on the correctness of the terminology. Overall, most systems translate 60%-70% of terminologies correctly. For terminology consistency, the most immediate outlier is $VARCO\text{-}MT_{TSSNMT}$, yielding impressive score of 0.971 on Chinese→English. Table 5 illustrates how even in the same document the terminology can be translated differently, which is undesired.

Source	Die Krankheit entwickelt sich bei Kindern und jungen Erwachsenen und folgt dem Muster der Blaschko-Linie.
MT	The condition develops during childhood and adolescence and follows the pattern of the blaschko line.
	...
Source	Ungefähr 95% aller Personen, die M. leprae ausgesetzt sind, entwickeln die Krankheit nicht.
MT	About 95% of all individuals exposed to M. leprae do not develop the disease.

Table 5: Example of term inconsistency (*Krankheit* → *disease, condition*) within the same document.

5.3 Terminology Utility (Tables 7 and 8)

Previous investigations into the general translation and terminology translation quality did not reveal many differences between the systems. We now focus on the usefulness of the additional information and show the difference between *Base* and either *Proper* or *Random* terminology dictionaries in Table 7. Notably *AdaptTerm* and *Lingua Custodia* improve the most from their *Base* version. With an exception of *OPUS-CAT*, both ChrF and COMET improves across all metrics when given any of the two dictionaries. This challenges the notion that the additional information supplied to the MT system needs to be terminology while in fact it can

System	Terminology Consistency De→En	En→Cs	Zh→En
AdaptTerm	0.617	0.753	0.750
Lingua Custodia	0.602	0.766	0.696
OPUS-CAT	0.661★	0.808★	0.293
$\text{UEDIN}_{\text{LLM}}$	0.588	0.741	0.713
$\text{UEDIN}_{\text{Tag}}$	0.606	0.750	0.755
$\text{UEDIN}_{\text{Twoshot}}$	0.574	0.737	0.622
BJTU-LB			0.764
$\text{VARCO-MT}_{\text{TSSNMT}}$			0.971
$\text{VARCO-MT}_{\text{ForceGen}}$			0.773★
Huawei	0.788	0.603	0.562

System	Terminology Success Rate De→En	En→Cs	Zh→En
AdaptTerm	0.591	0.577	0.785
Lingua Custodia	0.632	0.640	0.774
OPUS-CAT	0.948★	0.932★	0.133
$\text{UEDIN}_{\text{LLM}}$	0.557	0.594	0.750
$\text{UEDIN}_{\text{Tag}}$	0.532	0.584	0.765
$\text{UEDIN}_{\text{Twoshot}}$	0.560	0.498	0.452
$\text{VARCO-MT}_{\text{TSSNMT}}$			0.779
$\text{VARCO-MT}_{\text{ForceGen}}$			0.793★
BJTU-LB			0.759
Huawei	0.690	0.455	0.529

Table 6: Averages of Terminology Consistency and Terminology Success Rate with *Proper* terminology dictionaries. The ★ marks best within each column (language) and metric.

be any information that leaks from the reference. Focusing on a particular language pair in Table 8, there seems to be weak effect of lower variance when terminology dictionaries are provided.

6 Related Work

Similar to the previously shared task on translation using terminologies (Alam et al., 2021), our terminology hints are mined semi-automatically. We also extend this line of work by contrasting random and proper terminologies. The focus on terminologies in translation is an important one. Both Zouhar et al. (2020) and Semenov and Bojar (2022) show that the ordering of the system diverges when comparing performance on terminologies versus general performance.

Constrained Decoding. A simple paradigm for improving terminology translation is constrained decoding. Anderson et al. (2017) track constraint satisfaction using a finite-state machine. Hokamp and Liu (2017) reduce the time complexity to linear and Post and Vilar (2018) further improve on this.

Other approaches. Other than constrained decoding, several works have approached the problem by guiding the text generation model, including those that modify the token-level distribution using an external model (Stahlberg et al., 2017; Gulcehre et al., 2017; Chatterjee et al., 2017; Pascual et al., 2021), and those that incorporate constraints into the training process through additional annotations (Dinu et al., 2019; Bergmanis and Pinnis, 2021; Niehues, 2021, *inter alia*).

7 Conclusion

This iteration of machine translation with terminologies focused on evaluating the efficiency of using segment-level terminology dictionaries. I.e. it is not enough that the system performs well but it should also perform better when given this additional information. Indeed, the improvement between *Base* and *Proper* terminology enriched translations ranged across systems between 0 and 10 ChrF points. This helps isolate which terminology-enhancement methods are the most useful.

Limitations

The evaluation datasets are based on publicly-available data, which might have been leaked to the training of submitted systems, skewing the results. We further acknowledge that the comparisons in this work were not done using statistical testing.

Ethical Consideration

The work of both linguist working on the validation of GPT alignment was well-paid of around a twice to three times the minimal hourly wage in their respective countries. The annotated texts did not contain any sensitive or explicit passages.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical*

System	ChrF +Proper	ChrF +Random	$\text{COMET}_{22}^{\text{DA}}$ +Proper	$\text{COMET}_{22}^{\text{DA}}$ +Random	T. Consistency +Proper	T. Consistency +Random	T. Success Rate +Proper	T. Success Rate +Random
AdaptTerm	9.0	11.6	0.043	0.054	0.020	-0.010	0.3	0.338
Lingua Custodia	10.1	11.8	0.032	0.026	0.118	-0.016	0.402	0.369
OPUSCAT	10.2	9.2	0.031	0.043	0.055	0.187	0.345	0.247
$\text{UEDIN}_{\text{LLM}}$	6.4	7.5	0.011	0.017	0.027	0.018	0.214	0.157
$\text{UEDIN}_{\text{Tag}}$	5.4	6.5	0.010	0.013	0.055	0.009	0.218	0.127
$\text{UEDIN}_{\text{Twoshot}}$	6.9	5.9	0.029	0.012	0.045	-0.013	0.193	0.165
BJTU-LB †	2.5	0.8	0.015	0.007	0.058	0.049	0.252	-0.208
$\text{VARCO-MT}_{\text{TSSNMT}}$†	8.3	4.7	0.054	0.017	0.171	0.089	0.515	-0.041
$\text{VARCO-MT}_{\text{ForceGen}}$ †	3.4	0.9	0.019	0.003	0.166	0.021	0.529	-0.106
Huawei	0.2	0.9	-0.004	0.010	-0.010	-0.090	0.038	-0.113

Table 7: Average difference in each metric between the *Base* and added dictionary (*Proper* or *Random*). All numbers are averages across all languages except for † which is Chinese→English only.

System	$\text{COMET}_{22}^{\text{DA}}$ Zh→En Base	Proper	Random
AdaptTerm	$0.638_{\,0.142}$	$0.688_{\,0.109}$	$0.678_{\,0.104}$
Lingua Custodia	$0.476_{\,0.148}$	$0.609_{\,0.128}$	$0.528_{\,0.124}$
OPUSCAT	$0.521_{\,0.155}$	$0.557_{\,0.147}$	$0.624_{\,0.132}$
$\text{UEDIN}_{\text{LLM}}$	$0.750_{\,0.076}$	$0.757_{\,0.075}$	$0.753_{\,0.078}$
$\text{UEDIN}_{\text{Tag}}$	$0.747_{\,0.083}$	$0.757_{\,0.077}$	$0.747_{\,0.083}$
$\text{UEDIN}_{\text{Twoshot}}$	$0.572_{\,0.158}$	$0.650_{\,0.121}$	$0.596_{\,0.155}$
BJTU-LB	$0.736_{\,0.101}$	$0.751_{\,0.092}$	$0.743_{\,0.092}$
$\text{VARCO-MT}_{\text{TSSNMT}}$	$0.701_{\,0.145}$	$0.755_{\,0.138}$	$0.718_{\,0.135}$
$\text{VARCO-MT}_{\text{ForceGen}}$	$0.696_{\,0.094}$	$0.715_{\,0.091}$	$0.699_{\,0.095}$
Huawei	$0.679_{\,0.101}$	$0.666_{\,0.104}$	$0.709_{\,0.103}$

Table 8: Distribution of segment-level COMET scores on Chinese→English language direction (if available) between all three translation modes. Notation: mean $_{\text{var}}$.

Methods in Natural Language Processing, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2017. Leveraging bilingual terminology to improve machine translation in a CAT environment. *Natural Language Engineering*, 23(5):763–788.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with post-processing using constrained decoding and large language models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

M Teresa Cabré. 2010. Terminology and translation. *Handbook of translation studies*, 1:356–365.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020. Terminology-aware sentence mining for NMT domain adaptation: ADAPT's submission to the adap-MT 2020 English-to-Hindi AI translation shared task. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23, Patna, India. NLP Association of India (NLPAI).

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 2 (Short Papers), pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Md Mahfuz ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021. On the evaluation of machine translation for terminology consistency.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Jingshu Liu. 2023. Lingua custodia's participation at the WMT 2023 terminology shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023b. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Jan Niehues. 2021. Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. Association for Computational Linguistics.

Tommi Nieminen. 2023. Opus-cat terminology systems for the wmt23 terminology shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell, and Yeonsoo Lee. 2023. VARCO-MT: NCSOFT's WMT'23 terminology shared task submission. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rudolf Rosa and Vilém Zouhar. 2022. Czech and English abstracts of ÚFAL papers (2022-11-11). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain. Association for Computational Linguistics.

Chelo Vargas-Sierra. 2011. Translation-oriented terminology management and ICTs: present and future. *Interdisciplinarity and languages: Current Issues in Research, Teaching, Professional Applications and ICT.*, pages 45–64.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. WMT20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*, pages 371–380, Online. Association for Computational Linguistics.

Identify and annotate all terminology entities (consider only consecutive words) from source sentence and match them with the counterpart in the translated sentence.

____

Source 'en': after blowing your nose, coughing or sneezing.
Translation 'fr': après s'être mouché ou avoir toussé/éternué.
Annotation: {'en': 'coughing', 'fr': 'toussé'}, {'en': 'sneezing', 'fr': 'éternué'}

____

Source 'zh': 仙羽郡，武宗学府，后山林中，一个身披宽松武袍的削瘦少年，双盘下蹲，舌尖抵住牙齿，全身力量集中于左右两拳，轰打人粗大树。
Translation 'en': In mountainous forest behind Xianyu prefecture , martial arts training institute , there was thin young man wearing loose and comfortable martial artist robe . In the lotus position with his tongue against his teeth, he focused all his strength into both his fists and pummeled huge tree.
Annotation: {'zh':' 仙羽郡','en': 'Xianyu prefecture'}, {'zh':' 武宗学府','en': 'a martial arts training institute'}

____

Source 'en': According to Statistics Austria's current estimate from April 2015, expenditure for research and development carried out in Austria in 2015 is projected to grow nominally by around €271.36 million or 2.76% compared to 2014, thereby exceeding the €10 billion threshold for the first time (€10.10 billion).
Translation 'de': Gemäß der aktuellen Globalschätzung der Statistik Austria vom April 2015 werden die gesamten Ausgaben für Forschung und Entwicklung in Österreich 2015 voraussichtlich gegenüber dem Jahr 2014 um rd.271,36 Mio. €bzw. 2,76% nominell wachsen und damit erstmals die 10 Mrd. €-Schwelle überschreiten (10,10 Mrd. €).
Annotation: {'en': 'expenditure', 'de': 'gesamten Ausgaben'}, {'en': 'research and development', 'de': 'Forschung und Entwicklung'}, {'en': 'threshold', 'de': '-Schwelle'}

____

Source 'cs': Podle ředitele Institutu veřejné správy Filipa Hrůzy si pořadatelé nyní musí vyhodnotit, jestli je pro Brno závod výhodný.
Translation 'en': According to the Head of the Public Administration Institute, Filip Hrůza, the organizers must now assess whether the race brings benefits to Brno.
Annotation: {'en': 'Public Administration Institute', 'cs': 'Institutu veřejné správy'}, {'en': 'race', 'cs': 'závod'}

____

Source '**{source_lang}**': **{source_segment}**
Translation '**{target_lang}**': **{translated_segment}**
Annotation:

Prompt 2: The prompt for collecting aligned terminology with GPT-4. **Bolded** text is replaced with current segment.